Quantifying the extent to which index event biases influence large genetic association studies

HMG Advance Access published December 30, 2016

Title: Quantifying the extent to which index event biases influence large genetic association studies

Author list:

Hanieh Yaghootkar,[1] Michael P. Bancks,[2] Sam E. Jones,[1] Aaron McDaid,[3,4] Robin Beaumont,[1] Louise Donnelly,[5] Andrew R. Wood,[1] Archie Campbell,[6] Jessica Tyrrell,[1] Lynne J. Hocking,[7] Marcus A. Tuke,[1] Katherine S. Ruth,[1] Ewan R. Pearson,[5] Anna Murray,[1] Rachel M. Freathy,[1] Patricia B. Munroe,[8,9] Caroline Hayward,[10] Colin Palmer,[5] Michael N. Weedon,[1] James S. Pankow,[2] Timothy M. Frayling,[1,&,*] Zoltán Kutalik[3,4,&]

Authors' affiliations:

[1] Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, U.K.
[2] Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
[3] Institute of Social and Preventive Medicine, Lausanne University Hospital, 1010, Switzerland.
[4] Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland.
[5] Division of Cardiovascular & Diabetes Medicine, Medical Research Institute, University of Dundee, Dundee, Scotland, U.K.
[6] Generation Scotland, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, U.K.
[7] Institute of Medical Sciences, University of Aberdeen, Aberdeen, U.K.
[8] Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, U.K.
[9] NIHR Barts Cardiovascular Biomedical Research Unit, Barts and The London School of Medicine, Queen Mary University of London, London, U.K.
[10] Generation Scotland, MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, U.K.
([&]) These authors jointly supervised this work.
(*) Corresponding author

Abstract

As genetic association studies increase in size to 100,000s of individuals, subtle biases may influence conclusions. One possible bias is "index event bias" (IEB) that appears due to the stratification by, or enrichment for, disease status when testing associations between genetic variants and a disease-associated trait. We aimed to test the extent to which IEB influences some known trait associations in a range of study designs and provide a statistical framework for assessing future associations. Analysing data from 113,203 non-diabetic UK Biobank participants, we observed three (near *TCF7L2, CDKN2AB* and *CDKAL1*) overestimated (BMI-decreasing) and one (near *MTNR1B*) underestimated (BMI-increasing) associations among 11 type 2 diabetes risk alleles (at $P<0.05$). IEB became even stronger when we tested a type 2 diabetes genetic risk score composed of these 11 variants (-0.010 SDs BMI per allele, $P=5x10^{-4}$), which was confirmed in four additional independent studies. Similar results emerged when examining the effect of blood pressure increasing alleles on BMI in normotensive UK Biobank samples. Furthermore, we demonstrated that, under realistic scenarios, common disease alleles would become associated at $p<5x10^{-8}$ with disease-related traits through IEB alone, if disease prevalence in the sample differs appreciably from the background population prevalence. For example, some hypertension and type 2 diabetes alleles will be associated with BMI in sample sizes of >500,000 if the prevalence of those diseases differs by >10% from the background population. In conclusion, IEB may result in false positive or negative genetic associations in very large studies stratified or strongly enriched for/against disease cases.

Introduction

Genome wide association studies (GWAS) in increasingly large sample sizes have resulted in the identification of many 1000s of genetic variants associated with various common diseases(1-3). We assume that the results from GWAS are robust to potential confounding factors because genetic variants are inherited randomly and not influenced by disease processes throughout life. These assumptions tend to hold provided population substructure is accounted for using standard methods(4-6). Researchers may also run genetic association studies in stratified samples (1, 2) to reduce the non-genetic variation in the trait and improve statistical power to detect variants not acting through a disease mechanism. However, performing an association test between a genetic variant and a continuous trait in a sample that is stratified, depleted or enriched for a disease outcome (collider) that is associated with both the genetic variant and the trait can result in paradoxical observations (**Figure 1**). Such bias is termed index event bias (IEB) or collider-stratification bias(7).

An example of an index event biased association is that between type 2 diabetes risk alleles and lower BMI observed within diabetes case or control strata(8-10). In these examples the strata are enriched (type 2 diabetes cases) or depleted (type 2 diabetes controls) by disease status. Bias occurs because non-diabetic individuals with a diabetes protective allele are able to remain normo-glycaemic at higher BMIs than individuals without the protective allele, whilst individuals with a risk allele will tend to develop diabetes at lower BMIs compared to those without the risk allele. Considering this type of bias is very important because many large meta-analytic studies often perform GWAS analyses of traits in samples stratified by, enriched or depleted for disease status. Such bias can have different impacts on genetic associations and their interpretation, for example (a) in the case of true (positive) pleiotropy, the effect of the gene variant on the trait may be masked/reduced by IEB (false negative

finding); (b) in the case of no pleiotropy, a false (and often counterintuitive) association between gene variant and trait can be observed (false positive finding); (c) using a genetic variant as an instrumental variable in a Mendelian randomization analysis could result in false inferences about causal relationships between the disease and a correlated trait if estimates of that variants' effects on the trait are biased.

A research area closely related to IEB is secondary trait analysis in case-control sample design. Several approaches have been proposed in these settings to estimate the true effect between a risk factor and a genetic variant in the general population sample(11-16). Most of these methods yield unbiased estimators that are robust to model misspecifications and work for a wide range of noise distributions. In this paper we applied a retrospective likelihood maximization method (16), implemented in the R package SPREG, to correct for IEB for real data. Note, however, these methods provide only effect size estimation, but no insight how the bias explicitly depends on parameters such as (a) the extent of case-enrichment/depletion; (b) the strength of associations between the SNP, the risk factor and the collider; and (c) SNPs' allele frequency.

In this study, we tested the extent to which large genetic association studies may be impacted by IEB due to inadvertent sample selection leading to enrichment or depletion of disease cases. We first provided the statistical framework for quantifying IEB in a study, then used a combination of real and simulated data to: (a) identify and quantify real examples of IEB, including a single large study (120,000 individuals from the UK Biobank) and a meta-analysis of independent studies and (b) demonstrate that IEBs can occur in most types of genetic association study designs, e.g. 1:1 case-control designs to case only and control only studies, case or control enriched studies and when case status is used as a covariate. The

analytical formula is implemented in an R package and available for download from http://wp.unil.ch/sgg/files/2016/01/IndexEventBias.zip.

Material and Methods

Model formulation

We assumed a liability scale disease (probit) model, where a genetic variant (G) and a continuous risk factor (X) linearly influences the development of a disease (Y) on the liability scale. For simplicity the genetic marker was modeled by a binomially distributed random variable $G \sim B(2, q)$, where q is the allele frequency of the genetic variant in the general population. The second trait, from now on referred to as the "continuous risk factor" (X) was allowed to be linearly associated with G and assumed to follow a normal distribution $(X \,|G = g) \sim N(\gamma(g - 2q), 1 - 2\gamma^2 q(1 - q))$. In other words, $X = \gamma(g - 2q) + \varepsilon$, where ε is normally distributed with variance $1 - 2\gamma^2 q(1 - q)$. The disease status is then modeled as

$$Y = \begin{cases} 1 \; if \; Z > z_0 \\ 0 \; if \; Z \leq z_0 \end{cases} \text{ where } z_0 = \Phi^{-1}(1 - \pi_0)$$

with

$$(Z|X = x, G = g) \sim N(\alpha * x + \beta * (g - 2q), \sigma^2),$$

where π_0 is the disease prevalence in the general population; α is the true effect size on the liability scale of the continuous risk factor (X) on the disease outcome (Y); β is the (risk-factor independent) effect of the genetic variant (G) on the disease outcome (Y); and $\sigma^2 = 1 - \alpha^2 - 2\beta^2 q(1 - q)$.

Link between liability scale and logistic models

For simplicity, we derived the explicit analytical formula for IEB estimation for the liability scale model. This however does not prevent its applicability to parameters derived from the logistic regression model. By re-parameterizing the models one can reach indistinguishable

properties(17). Namely, we calculate the probability of (Y=0 | X, G) for data simulated from the logistic model and optimize the liability scale model parameters such that this probability is matched as close as possible for each simulated data point. We noticed that the models are indistinguishable when the parameter optimization is done for a simulated population with similar disease prevalence to the tested sample. The details of this procedure are given in the Appendix (section 2.3).

Analytical formula for IEB

The extent of IEB can be analytically derived in the case of the liability scale (probit) disease model. Assume that disease frequency in the general population is π_0, the allele frequencies of the genotype in pure control and pure case populations are q_0 and q_1, respectively. We denote the difference between these frequencies by δ_q. Let us consider now a sample with π frequency of cases. The allele frequency in this sample is

$$q_\pi = q_0 + \pi * (q_1 - q_0) = q_0 + \pi * \delta_q$$

IEB occurs when this disease prevalence differs from the general population prevalence (π_0). When $\pi > \pi_0$, we observe an enrichment of cases, while in case of $\pi < \pi_0$, we observe an enrichment of controls. In the Supplementary Text we derived the per-allele linear regression effect size of G on X in the general case, but here for simplicity we present the formula assuming $\gamma = 0$, i.e. that the true underlying effect of the genotype on the risk factor is zero. We observed that this simplification makes very little difference in practice (**Supplementary Figure 1**). By introducing the quantities

$$\sigma^2_{Z|G} = 1 - 2\beta^2 q(1-q)$$

$$w_j = \frac{z_0 - \beta(j - 2q)}{\sigma_{Z|G}}$$

we can express the expectation of the linear regression effect size estimate of G on X as

$$E[\widehat{\gamma_\pi}] \approx \frac{\alpha}{2q_{\boldsymbol{\pi}}(1-q_{\boldsymbol{\pi}})} \cdot \frac{\boldsymbol{\pi}-\pi_0}{\pi_0(1-\pi_0)} \left(\frac{2q(1-q)\varphi(w_1) + 2q^2\varphi(w_2) - 2\varphi(z_0)(\boldsymbol{\pi}\delta_q + q_0)}{\sigma_{Z|G}} \right)$$

where $\varphi()$ denotes the probability density function (pdf) of the standard Gaussian distribution. Note that when the disease prevalence matches the general population prevalence (i.e. $\pi = \pi_0$), the expected effect of G on X is unbiased. The formula also shows that the bias is a ratio of two expressions quadratic in the prevalence (π). For most settings the quadratic and linear terms of the denominator are small compared to that of the numerator, thus the bias as a function of the fraction of cases (π) closely resembles a simple parabolic function. It is worth noting that the coefficient of the quadratic term (in π) in the numerator $\left(-\frac{\varphi(z_0)\delta_q}{\sigma_{Z|G}\,\pi_0(1-\pi_0)}\alpha\right)$ has the opposite sign compared to α, meaning that when X is a risk factor for the disease the function is a downward looking parabola. This explains why disease risk alleles can show spurious (and counterintuitive) protective effect on traits positively correlated with the disease. We also derived a formula for the case when we do not assume that the true effect is zero; i.e. when there is a true pleiotropy. The full derivation of the formula can be found in the Appendix (Section 2).

Note that this formula assumes that the true parameters, $\alpha, \beta, q_0, q_1, q, \pi_0$ are known. Hence, its primary purpose is not to estimate the bias from data, but to reveal the intricate relationship between the true underlying model parameters and the resulting IEB.

We extended the formula to situations when not only the sample is enriched or depleted for a disease, but also when in addition the continuous risk factor is corrected for the disease status:

$$E[\widehat{\gamma_\pi}^{(C)}] = \frac{\gamma_\pi q_\pi(1-q_\pi) - \delta_q\, \alpha\, \varphi(z_0) \dfrac{\pi(1-\pi)}{\pi_0(1-\pi_0)}}{q_\pi(1-q_\pi) - 2\,\pi(1-\pi)\delta_q{}^2}$$

The full derivation of this formula is given in the Appendix (Section 3). One can observe that correcting for disease status yields biased estimates even when the study disease prevalence agrees with the general population prevalence. In this situation we showed that this formula simplifies to that of Aschard(18).

These analytical formulae for IEB allow us to quantify genetic effects (estimated from real data) for IEB. The key quantities necessary for the formulae are: (i) the allele frequency of the genetic variant in control (q_0) and disease (q_1) populations; (ii) the association effect of the SNP on the disease status (β); (iii) the effect of the continuous risk factor on the disease status (α); (iv) the disease prevalence in the study population (π) and (v) the general population disease prevalence (π_0). Estimating these quantities is out of the scope of this paper. We use the formula to show the extent of the bias for various realistic parameter settings informed from large GWAS data.

Data simulation to confirm the analytical formula

To investigate how closely the analytical formula recapitulates true IEB, we simulated data to create different hypothetical scenarios similar to data used in genetic studies (**Supplementary Figure 2**). We simulated binomially distributed SNP data (G), a normally distributed continuous risk factor (X) and binary disease status using the liability threshold model described above. The minor allele frequency (MAF) of the genetic marker was explored in the range of 0.02, 0.05, 0.1, 0.3 and 0.5; disease prevalence in the general population was set to 1%, 5% and 10%. The effect of the continuous risk factor (X) on the disease outcome (Y) was varied in a range equivalent to ORs of 1.10 to 4 (1.10, 1.25, 1.50,

1.75, 2, 3 and 4) per 1 standard deviation (SD) increase in X. The effect (β) of the genetic variant (G) on the disease outcome (Y) was explored for a range equivalent to odds ratios (OR) of 1.05, 1.2, 1.3, 1.5, 1.8 and 2. We simulated 101 settings with 100,000 total samples, for case-control ratios ranging from 0 to 100%.

IEB in real data

To identify and quantify real examples of IEB, we tested how IEB occurs in genetic studies in 2 different scenarios using different disease outcomes and genetic variants.

(1) We tested whether or not type 2 diabetes risk alleles, acting predominantly through insulin secretion, have a paradoxical (opposite association) effect on BMI, a strong continuous risk factor for type 2 diabetes. We selected 11 SNPs associated with type 2 diabetes that have known robust associations with insulin secretion (19), the intermediate trait most relevant to diabetes risk (19) (**Supplementary Table 1**). On purpose we avoided the use of insulin resistance variants, which could operate through adiposity. We analysed the 11 SNPs separately and as a genetic risk score (GRS) in two study types: (i) a single very large population based study: 120,286 individuals from the first release of genetic data from the UK Biobank study(20) and (ii) a meta-analysis of 4 independent studies: EXTEND(21) (N = 5,097), GoDARTS(22) (N = 7,128), Generation Scotland Scottish Family Health Study (GS:SFHS)(23) (N = 8,195) and ARIC(24) (N = 9,324) with a range of study designs and diabetes status available (**Supplementary Table 2**). We tested the association between individual SNPs and the 11-SNP insulin secretion GRS and BMI in all samples, in all samples adjusted for type 2 diabetes status, in diabetic cases only and in controls only. We defined type 2 diabetes cases as individuals who had: (1) HbA1c >6.4% and/or fasting glucose >7 mmol/L, (2) age at diagnosis >35 and <70 years; (3) no need for insulin treatment

within 1 year of diagnosis (except in ARIC). We defined controls as individuals who did not meet any of these criteria and were not diagnosed with any other types of diabetes. We additionally tested the association between the *CCND2* type 2 diabetes protective allele and BMI as the allele was recently shown to be associated with higher BMI.

(2) We tested whether or not 29 SNPs robustly associated with systolic blood pressure have a paradoxical (opposite) effect on BMI, a continuous risk factor for high blood pressure(25). 26 of the 29 variants were reliably imputed in the UK Biobank and we excluded the variant near *SLC39A8* from the GRS as this variant is directly associated with several traits including BMI(26) and HDL-cholesterol(27) levels (**Supplementary Table 3**). We analysed the remaining 25 SNPs individually and as a GRS in two study types: (i) a single very large study: 120,286 individuals from the first release of genetic data from the UK Biobank study(20) and (ii) a meta-analysis of 4 independent studies with blood pressure available: GoDARTS(22) (N = 6643), Generation Scotland Scottish Family Health Study (GS:SFHS)(23) (N = 8195), ARIC(24) (N= 9,290) and BRIGHT(28) (N = 1808) (**Supplementary Table 2**). We tested the association between the individual SNPs and the 25-SNP blood pressure GRS and BMI in all samples, in all samples adjusted for hypertension status, in hypertensive cases only and in normotensive controls only. We defined hypertensive cases as individuals with systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or report use of anti-hypertensive medications. We defined normotensive controls as individuals with systolic and diastolic blood pressure below these thresholds, and not on medications.

Statistical analysis in real data

In the relevant studies, we corrected BMI for age and sex and other covariates (for the UK Biobank study this included five within UK genetic principal components, genotyping

platform, study center in the UK Biobank study). Residuals were inverse-normal transformed. In each study, we generated genetic risk scores (GRS) by calculating the number of disease risk alleles carried by each individual. We then combined the association results using fixed-effects inverse variance-weighted meta-analysis. To account for IEB we applied the state-of-the-art method of Lin and Zeng(16) (implemented in the software SPREG). This program needs an estimate of the general population prevalence of the examined diseases. Hence, we derived an estimate for type 2 diabetes and hypertension prevalence for a general UK sub-population that has the same joint age- and sex-distribution as the UK Biobank sample. For this we used sex- and age-group-specific prevalence values from the IDF Atlas (29) (10 year bins) for type 2 diabetes and from the NIH Health Survey for England 2011 [http://digital.nhs.uk/catalogue/PUB09300/HSE2011-Ch3-Hypertension.pdf] (10 year bins) for hypertension. We then weighted these prevalence values with the proportion of UK Biobank participants that fell into each stratum. This yielded prevalence estimates of 10.15% for type 2 diabetes and 38.43% for hypertension.

Results

IEB occurs in a range of different genetic association study designs – theory and simulated data.

Our results from simulated data and theory provided examples of IEBs where the direction of the association between the genetic risk allele and the risk factor depends on the proportion of cases and controls in the study. In all scenarios involving a disease risk allele with no real association to the intermediate risk factor, we observed U-shaped artefactual effect estimates between the disease risk allele and the risk factor as the proportion of cases moved from 0% to 100%: first, in a control only situation, IEB occurred where the disease risk alleles were associated with lower values of the risk factor, there was then no association when the proportion of cases represented exactly the background population, then an association between disease risk alleles and higher values of the risk factor, and then back to no association and finally an association between disease risk alleles and lower values of the risk factor in case only scenarios (**Figure 2** [our theoretical formula]; **Supplementary Figure 3** [simulated data]). The extent of the bias is stronger in case only compared to control only scenarios when the disease frequency is less than 50% (as with most diseases). In the examples in **Figure 2** (and **Supplementary Figure 3**), we modeled a disease risk allele and a protective allele with properties similar to those of the type 2 diabetes alleles at *TCF7L2* and *CCND2* respectively. We observed spurious associations between the disease alleles and lower and higher values of the continuous risk factor, depending on the proportion of cases and despite the lack of a genuine association between the genetic risk allele and the risk factor. When the examined study population matches the underlying general population in terms of disease prevalence (5% in case of our example), no bias is observed (**Figure 2**). It has been shown that for many scenarios a simple regression including the disease status as

covariate alleviates the bias(15), but such correction is clearly of no use in case-only, and control-only designs and it introduces bias even for samples representative of the general population. Note that in our settings the bias was not resolved, but often exacerbated by correcting for disease status (**Figure 2**).

Differences in prevalence for Type 2 diabetes and hypertension in the UK Biobank and general UK population

Index event bias arises due to differences in the (collider) disease prevalence in the study population and the matched general population. Hence, it is crucial to derived accurate estimates for type 2 diabetes and hypertension prevalence for a general UK sub-population that has the same joint age- and sex-distribution as the UK Biobank. Using data from the IDF Atlas (29) and the NIH Health Survey for England, we estimated that 10.15% and 38.43% of a sex- and age-matched sub-population of the UK would be diabetic and hypertensive, respectively (see Methods). These values clearly differ from the prevalence of type 2 diabetes (3.4%) and hypertension (55.2%) observed in the UK Biobank.

Individual alleles and genetic risk scores associated with higher risk of type 2 diabetes were associated with lower BMI in real data

A relatively high ability to secrete insulin may lead to a relative protection from type 2 diabetes but may also lead to higher BMI because insulin has anabolic properties. Studies may therefore wish to use common variants associated with insulin secretion to test the role of insulin secretion on BMI. However, there may be a complex relationship because higher BMI increases diabetes risk. IEB will add to the complexity of interpreting potential overlap of genetic associations for these phenotypes. We tested 11 variants associated with type 2 diabetes through an insulin secretion mechanism, for potentially spurious associations with lower levels of the continuous risk factor for type 2 diabetes, BMI. Details of how these

variants were associated with type 2 diabetes in UK Biobank and 4 additional studies are given in **Supplementary Table 4**. Using a total of 4,003 type 2 diabetes cases and 113,203 controls from the UK Biobank, two of the 11 variants were associated with lower BMI in all individuals (unstratified and unadjusted), three in controls only, three in cases only and five in all individuals when adjusted for type 2 diabetes status at $p<0.05$ (**Table 1**). When meta-analysing the UK Biobank and four additional studies in the same analysis design as a GWAS meta-analysis (all individuals together for population based studies, stratified by case control status for case control studies) two type 2 diabetes risk alleles were associated at $p<0.05$ with lower BMI (**Supplementary Table 5**). In the UK Biobank study, the effect sizes of type 2 diabetes risk alleles with lower BMI were consistent with IEB (**Supplementary Figure 4)**. In accordance with our formula, BMI "effect" size estimates were correlated with the effect estimates for type 2 diabetes; r = -0.85 (p=8E-4), -0.87 (p=4E-4) and -0.87 (p=5E-4) in controls, cases and all individuals (adjusted for type 2 diabetes status), respectively (**Supplementary Figure 4**).

We next reran the SNP-BMI associations, using a retrospective likelihood based method implemented in the statistical software SPREG, which accounts for IEB(16). Prior to this correction, the risk allele at *TCF7L2* was associated with lower BMI in all scenarios (the overall population as well as stratified and corrected data - **Table 1**). After correcting for IEB there was no evidence (at $p<0.005$; p-value corrected for multiple testing) for an association between *TCF7L2* and lower BMI (**Table 1**). In contrast, the type 2 diabetes risk allele at *MTNR1B* was the only allele associated with higher BMI in the overall population ($p = 0.02$); when accounting for IEB it was even more strongly associated (0.016 SD [0.007, 0.026], P=0.001) (**Table 1, Figure 3)**. The type 2 diabetes protective allele in *CCND2* (conferring the strongest effect on type 2 diabetes; 0.59 OR [0.48,0.73]; p = 1E-6) had the strongest effect

estimate on BMI (0.06 SD [0.03,0.09]; p = 0.0004), which became much weaker after correcting for IEB (0.003 SD [-0.029,0.035]; p = 0.9).

We next examined a genetic risk score (GRS) for type 2 diabetes. Details of how this GRS was associated with type 2 diabetes are given in **Supplementary Table 4**. In the UK Biobank study, the 11-SNP GRS was not associated with BMI when analysed in all samples combined (-0.004 SD per allele [-0.010,0.001]; p=0.1; N=119,688; **Table 2**). In contrast, the 11-SNP GRS associated with higher risk of type 2 diabetes was associated with lower BMI in all 3 of the following designs: (i) controls only (-0.010 SD per allele [-0.016,-0.005]; p=5E-4; N=113,203), (ii) in cases only (-0.036 SD per allele [-0.065,-0.007]; p=0.01; N=4,003) and (iii) in all individuals when adjusted for type 2 diabetes status (-0.011 SD per allele [-0.017,-0.006]; p=1E-4; N=117,206; **Table 2**). In the context of a Mendelian randomization analysis, these results could be misinterpreted as evidence for the biologically plausible hypothesis that lower insulin secretion leads to lower BMI. However the associations are consistent with IEB. Results from a meta-analysis of 4 additional studies (representing a scenario similar to that of many GWAS meta-analyses) were similar (**Table 2** and **Figure 4a**).

Individual alleles and genetic risk scores associated with higher risk of hypertension were associated with lower BMI

We next tested whether alleles associated with higher risk of hypertension were paradoxically associated with lower BMI, a continuous risk factor for hypertension, but with a weaker effect than that with type 2 diabetes. Such associations could be due to genuine pleiotropic effects of alleles on hypertension and lower BMI, or due to IEB, or a combination of the two. We tested 25 variants associated with blood pressure. Details of how these variants were associated with hypertension in UK Biobank and four additional studies are given in **Supplementary Table 6**. Using a total of 65,584 hypertension cases and 53,377 controls

from the UK Biobank, six of the 25 variants were associated with lower BMI in all individuals (unstratified and unadjusted), 10 in controls only, 10 in cases only and 12 in all individuals when adjusted for hypertension status (at p<0.05; **Table 3**). When meta-analysing the UK Biobank and four additional studies, in the same analysis design as a GWAS meta-analysis (all individuals together for population based studies, stratified by case control status for case control studies), eight hypertension risk alleles were associated at p<0.05 with lower BMI (**Supplementary Table 7**). The effect sizes of hypertension risk alleles with lower BMI were consistent with IEB. As with type 2 diabetes alleles, BMI "effect" sizes were correlated with the effect size on hypertension: r = -0.38 (p=0.06), -0.58 (p=0.002) and -0.55 (p=0.005) in controls, cases and all individuals (adjusted for hypertension status), respectively (**Supplementary Figure 5**).

Using SPREG, out of the 25 hypertension SNPs only *CYP17A1* was associated with lower BMI in the IEB corrected analysis (-0.028 [-0.044,-0.012]; P=3E-4) and Bonferroni correction for the number of SNPs tested (**Table 3**). Five other variants (those in or near *BAT2, CACNB2* (2 variants)*, CYP1A1,* and *SH2B3*) were associated with lower BMI at p<0.05 but did not persist after Bonferroni correction. Nevertheless, six variants reaching IEB-corrected nominally significant P-values is more than the ~1 expected by chance (enrichment P = 1.69E-4) and suggests variants in some of these genes have pleiotropic effects with alleles associated with lower BMI and higher risk of hypertension. Consistent with this evidence of pleiotropy, the variant in *SH2B3* is associated with multiple traits including those related to autoimmunity as well as metabolic traits(30-32).

We next considered a genetic risk score of hypertension SNPs. Details of how this GRS was associated with hypertension are given in **Supplementary Table 6**). In the UK Biobank study, the 25-SNP hypertension GRS was associated with lower BMI in all samples

combined (-0.014 SD per allele [-0.020,-0.008]; p=1E-6; N=119,688) and in all 3 of the following designs: (i) in controls only (-0.034 SD per allele [-0.043,-0.026]; p=2E-16; N=53,377), in cases only (-0.031 SD per allele [-0.039,-0.024]; p=4E-16; N=65,584) and in all samples when adjusted for hypertension status (-0.033 SD per allele [-0.038,-0.027]; p=7E-31; N=118,961; **Table 2**). Results from a meta-analysis of 4 studies (representing a GWAS meta-analysis) were similar (**Table 2** and **Figure 4b**).

Large sample sizes are necessary to observe false positive associations due to IEB

Our analytical formula enabled us to quantify the necessary sample size in order to observe a false positive association with a continuous risk factor due to IEB at any significance level alpha with for example, 80% power. The necessary sample size depends on five parameters: significance level, disease prevalence, strength of association between the genetic risk factor and the disease, strength of association between the continuous risk factor and the disease and frequency of the genetic risk allele. We fixed the continuous risk factor-disease association (OR=2.5 per SD) and tested two MAF scenarios (low, 2% and medium 30%) and two significance levels (0.05 and 5E-8). The remaining two parameters (SNP-disease association strength and disease prevalence) we varied freely and computed the minimal sample size necessary to detect a false association (**Figure 5**). For example, in analyses stratified by disease we would need 23,542 cases or 208,267 controls to detect a biased association at p-value $5x10^{-8}$ with a probability of 80% when the disease risk allele had a frequency of 30%, the disease prevalence was 10% and a 1 SD higher value of the continuous risk factor was associated with an odds ratio of 2.5 for the disease. This scenario is similar to that for the risk allele at *TCF7L2* and BMI(10). Notably, scenarios with SNP-disease OR=1.2 in 500,000 disease-free samples will yield (with 80% probability) false positive genome-wide significant SNP-risk factor association entirely due to IEB.

Discussion

Our analyses of real and simulated data showed that index event bias will often occur in genetic association studies but the extent depends on several factors. These factors include the association strength between the trait being analysed (that we termed the "continuous risk factor") and the disease, where the disease is over- or under-represented in the study population compared to the background population. The other factors are disease prevalence, sample size, and the effect size and minor allele frequency of the disease associated variant.

Our results go beyond those of previous studies examining index event biases in several ways. First we provide real examples of likely biased genetic associations in the context of studies of 100,000s of individuals, including those involving individual variants and combinations of variants. Second, we provide a formula for quantifying the bias as a function of key parameters even when only summary level data is provided. We also extended the work of Aschard *et al*(18), to test how the combination of correcting for disease status in disease-enriched or depleted samples can introduce biases.

Our results have important implications for all types of large genetic association studies, and are especially relevant given that analyses are now possible in 100,000s of individuals, and rarely will these samples be perfectly representative of the background populations – for example, even population based studies such as those of Decode, 23andMe and the UK Biobank are likely not truly representative of the background population in the prevalence of all disease outcomes. Our analyses of real and simulated data showed that the best study design to avoid index event biased associations is using all individuals from a population-based study with no adjustment for disease status. Bias is strongest in case only designs (assuming the disease frequency is <50%) but it is also observed in control only designs, or in analysis combining cases and controls and adjusting for disease status (the latter situation is

discussed in Aschard et al(18)). To better understand these observations, we derived an analytical formula that estimates IEB and confirmed its validity through simulation studies. Our formula indicates that this bias can be negative or positive depending on the proportion of cases and controls. The results indicate that this bias cannot be resolved by correcting for disease status as a covariate. If anything, such correction exacerbates biased effect estimates. Our results also indicate that the impact of IEB is substantially larger than the bias caused by improper covariate correction in a disease-representative population (described by Aschard *et al*(18)).

In a meta-analysis it may be difficult to assess the extent to which IEB is contributing to an association - most existing large scale genetic association studies are mixtures of all types of study designs, and, for studies of continuous traits (such as BMI, lipid levels or blood pressure) disease cases and controls are often analysed as separate strata before meta-analysis with population-based studies, which themselves could be over or under represented with disease cases.

Our settings for the analytical formula were limited to a liability scale disease model and normal linear regression applied for the risk trait. By model re-parameterization we extended it to the logistic disease model and through simulations we saw that it works equally well (data not shown). These are the most often used models in meta-analytic GWAS studies; hence we believe that our findings are extremely relevant for almost all GWAS analysis.

Whilst index event biases are likely to exist in many studies, for associations modelling individual variants the bias is unlikely to cause false positive or false negative associations unless sample sizes are very large or stratified, strongly depleted of, or enriched for, disease cases. For example, we tested known common type 2 diabetes variants for association with a strong risk factor for type 2 diabetes, BMI, in 119,688 UK Biobank individuals (including

4003 type 2 diabetes cases), but only the most strongly associated diabetes variant, that in *TCF7L2* (odds ratio ~1.4), was associated with lower BMI at $p<0.01$ in all individuals. The association between the variant in *TCF7L2* and BMI has been the subject of several previous papers(1, 8, 33) and was most recently noted as showing stronger effect estimates in case control studies(1). Here we found only a trend, but no clear statistical evidence that the *TCF7L2* variant is associated with BMI - the associations we did see could be explained by IEB due to a likely depletion of cases in the UK Biobank study compared to the background population.

The derived analytical formula can serve as guidance for the expected bias in genome-wide association studies, where only summary level data is known where – to our knowledge – no other method is applicable. The state-of-the-art tool, SPREG, computed the corrected effect in ~2 min per SNP, which renders such methods infeasible for large population cohorts with genome-wide genotype data, as it would take >10 CPU years to apply for millions of markers.

Accounting for IEB strengthened associations for several individual variants with good prior evidence for pleiotropic effects on the disease and continuous risk factor. For example, the type 2 diabetes risk allele at *MTN1RB* was associated in the UK Biobank with higher BMI and this result strengthened on correction for IEB – results from previous studies, particularly those that were not population-based, may have been biased towards the null. This variant has one of the strongest effects on fasting glucose levels in individuals without diabetes and may predispose to higher BMI through higher insulin secretion. The hypertension risk allele at *CYP17A1* was previously associated with lower BMI(1), and we show here that this is a likely pleiotropic effect.

Studies examining the joint effect of multiple variants will be more prone to index event biases than those of single variants. Studies prone to miss-interpretation could include gene-based tests, and tests of the cumulative effect of variants when using a genetic risk score. For example, Mendelian randomization studies often use genetic risk scores as instruments to test causality of an associated trait with a disease(34-37). Stratified Mendelian randomization is recommended when the exposure is binary (e.g. smoking) and hence a causal effect should be seen only in the exposed stratum(38). Such stratification in some cases could introduce IEB in the causal estimation(38). Failing to account for IEB could lead to false conclusions about causality. We explored this potential source of bias using the UK Biobank to assess whether or not a genetic risk score of insulin secretion (represented by 11 variants associated with type 2 diabetes through insulin secretion mechanisms) was associated with BMI. An association between a genetic risk score for poorer insulin secretion and lower BMI could indicate that insulin secretion causally alters BMI, a plausible hypothesis given that insulin treatment increases BMI in diabetes(39). However, IEB would also result in an association between a genetic risk score for poorer insulin secretion (type 2 diabetes risk alleles) and lower BMI. Whilst we cannot disentangle IEB from a genuine pleiotropic effect IEB is the more likely explanation given the gradient of stronger effects in cases compared to controls compared to all individuals (**Supplementary Figure 6a**). Similar analysis for hypertension provided evidence that SNPs associated with higher blood pressure are also associated with lower BMI (**Supplementary Figure 6b**).

In summary, as genetic association studies reach sizes of 100,000s of individuals, analyses will be prone to misinterpreting results if they do not account for index event biases. However, we have provided the statistical framework and its software implementation for quantifying and correcting for these biases under reasonable assumptions. Because IEB is dependent on many different factors it may occur in a variety of situations and can cause false

positive and false negative results. Researchers should be particularly vigilant when either or all of the following occur: i) a known disease variant is associated with a known risk factor for that disease, ii) the association is in the opposite direction to that seen observationally between disease and risk factor, iii) the sample size or effect size of the allele on the disease are particularly large. As a rule of thumb, we suggest that, for single variants, people should take extra care when disease odds ratios are >1.4 and sample sizes are >100,000.

Acknowledgments. This research has been conducted using the UK Biobank Resource. The authors thank University of Exeter Medical School. EXTEND data were provided by the Peninsula Research Bank, part of the NIHR Exeter Clinical Research Facility. P.B.M. wishes to acknowledge support from the NIHR Cardiovascular Biomedical Research Unit at Barts and The London, Queen Mary University of London, UK. We are grateful to all the participants who took part in the GS:SFHS study, to the general practitioners, to the Scottish School of Primary Care for their help in recruiting the participants, and to the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. The Wellcome Trust provides support for Wellcome Trust United Kingdom Type 2 Diabetes Case Control Collection (GoDARTS) and informatics support is provided by the Chief Scientist Office. The Atherosclerosis Risk in Communities Study (ARIC) is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

H.Y., A.R.W. and T.M.F. are supported by the European Research Council grant: 323195; SZ-245 50371-GLUCOSEGENES-FP7-IDEAS-ERC. S.E.J. is funded by the Medical Research Council (grant: MR/M005070/1). M.A.T., M.N.W. and A.M. are supported by the Wellcome Trust Institutional Strategic Support Award (WT097835MF). R.M.F. is a Sir

Henry Dale Fellow (Wellcome Trust and Royal Society grant: 104150/Z/14/Z). R.B. is funded by the Wellcome Trust and Royal Society grant: 104150/Z/14/Z. J.T. is funded by a Diabetes Research and Wellness Foundation Fellowship. Z.K. received financial support from the Leenaards Foundation, the Swiss Institute of Bioinformatics and the Swiss National Science Foundation (31003A-143914) and SystemsX.ch ((40)). The work of M.P.B was supported by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number T32HL007779. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. E.R.P. holds a WT New investigator award 102820/Z/13/Z.

We would like to thank George Davey Smith, Mark McCarthy and Joel Hirschhorn for helpful comments on the manuscript.

References

1 Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J. *et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197-206.

2 Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E. *et al.* (2015) New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, **518**, 187-196.

3 Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, **46**, 1173-1186.

4 Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R. *et al.* (2008) Genes mirror geography within Europe. *Nature*, **456**, 98-101.

5 Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, **42**, 348-354.

6 Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, **44**, 821-824.

7 Choi, H.K., Nguyen, U.S., Niu, J., Danaei, G. and Zhang, Y. (2014) Selection bias in rheumatic disease research. *Nat Rev Rheumatol*, **10**, 403-412.

8 Helgason, A., Palsson, S., Thorleifsson, G., Grant, S.F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I. *et al.* (2007) Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet*, **39**, 218-225.

9 Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A. *et al.* (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet*, **46**, 294-298.

10 Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661-678.

11 Monsees, G.M., Tamimi, R.M. and Kraft, P. (2009) Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol*, **33**, 717-728.

12 Tapsoba Jde, D., Kooperberg, C., Reiner, A., Wang, C.Y. and Dai, J.Y. (2014) Robust estimation for secondary trait association in case-control genetic studies. *Am J Epidemiol*, **179**, 1264-1272.

13 Lin, D.Y., Zeng, D. and Tang, Z.Z. (2013) Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proc Natl Acad Sci U S A*, **110**, 12247-12252.

14 Song, X., Ionita-Laza, I., Liu, M., Reibman, J. and We, Y. (2016) A General and Robust Framework for Secondary Traits Analysis. *Genetics*, **202**, 1329-1343.

15 Tchetgen Tchetgen, E.J. (2014) A general regression framework for a secondary outcome in case-control studies. *Biostatistics*, **15**, 117-128.

16 Lin, D.Y. and Zeng, D. (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol*, **33**, 256-265.

17 Sham, P.C., Walters, E.E., Neale, M.C., Heath, A.C., MacLean, C.J. and Kendler, K.S. (1994) Logistic regression analysis of twin data: estimation of parameters of the multifactorial liability-threshold model. *Behav Genet*, **24**, 229-238.

18 Aschard, H., Vilhjalmsson, B.J., Joshi, A.D., Price, A.L. and Kraft, P. (2015) Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet*, **96**, 329-339.
19 Dimas, A.S., Lagou, V., Barker, A., Knowles, J.W., Magi, R., Hivert, M.F., Benazzo, A., Rybin, D., Jackson, A.U., Stringham, H.M. *et al.* (2014) Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes*, **63**, 2158-2171.
20 Collins, R. (2012) What makes UK Biobank special? *Lancet*, **379**, 1173-1174.
21 http://exeter.crf.nihr.ac.uk/node/155. in press.
22 Morris, A.D., Boyle, D.I., McMahon, A.D., Greene, S.A., MacDonald, T.M. and Newton, R.W. (1997) Adherence to insulin treatment, glycaemic control, and ketoacidosis in insulin-dependent diabetes mellitus. The DARTS/MEMO Collaboration. Diabetes Audit and Research in Tayside Scotland. Medicines Monitoring Unit. *Lancet*, **350**, 1505-1510.
23 Smith, B.H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S.M., Deary, I.J., Macintyre, D.J., Campbell, H., McGilchrist, M. *et al.* (2013) Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol*, **42**, 689-700.
24 The ARIC investigators. (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. . *Am J Epidemiol*, **129**, 687-702.
25 Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., Hwang, S.J. *et al.* (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**, 103-109.
26 Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Magi, R. *et al.* (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, **42**, 937-948.
27 Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat Genet*, **45**, 1274-1283.
28 Caulfield, M., Munroe, P., Pembroke, J., Samani, N., Dominiczak, A., Brown, M., Benjamin, N., Webster, J., Ratcliffe, P., O'Shea, S. *et al.* (2003) Genome-wide mapping of human loci for essential hypertension. *Lancet*, **361**, 2118-2123.
29 , in press.
30 Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y. *et al.* (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature*, **480**, 201-208.
31 Eriksson, N., Tung, J.Y., Kiefer, A.K., Hinds, D.A., Francke, U., Mountain, J.L. and Do, C.B. (2012) Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One*, **7**, e34442.
32 Gudbjartsson, D.F., Bjornsdottir, U.S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G.M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H. *et al.* (2009) Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat Genet*, **41**, 342-347.
33 Timpson, N.J., Lindgren, C.M., Weedon, M.N., Randall, J., Ouwehand, W.H., Strachan, D.P., Rayner, N.W., Walker, M., Hitman, G.A., Doney, A.S. *et al.* (2009) Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes*, **58**, 505-510.

34 Yaghootkar, H., Lamina, C., Scott, R.A., Dastani, Z., Hivert, M.F., Warren, L.L., Stancakova, A., Buxbaum, S.G., Lyytikainen, L.P., Henneman, P. *et al.* (2013) Mendelian randomization studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. *Diabetes*, **62**, 3589-3598.
35 Fall, T., Xie, W., Poon, W., Yaghootkar, H., Magi, R., Knowles, J.W., Lyssenko, V., Weedon, M., Frayling, T.M. and Ingelsson, E. (2015) Using Genetic Variants to Assess the Relationship Between Circulating Lipids and Type 2 Diabetes. *Diabetes*, **64**, 2676-2684.
36 Holmes, M.V., Dale, C.E., Zuccolo, L., Silverwood, R.J., Guo, Y., Ye, Z., Prieto-Merino, D., Dehghan, A., Trompet, S., Wong, A. *et al.* (2014) Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ*, **349**, g4164.
37 Nuesch, E., Dale, C., Palmer, T.M., White, J., Keating, B.J., van Iperen, E.P., Goel, A., Padmanabhan, S., Asselbergs, F.W., Verschuren, W.M. *et al.* (2015) Adult height, coronary heart disease and stroke: a multi-locus Mendelian randomization meta-analysis. *Int J Epidemiol*, in press.
38 Gage, S.H., Davey Smith, G., Ware, J.J., Flint, J. and Munafo, M.R. (2016) G = E: What GWAS Can Tell Us about the Environment. *PLoS Genet*, **12**, e1005765.
39 Larger, E., Rufat, P., Dubois-Laforgue, D. and Ledoux, S. (2001) [Insulin and weight gain: myth or reality?]. *Diabetes Metab*, **27**, S23-27.
40 Berkson, J. (1946) Limitations of the application of fourfold table analysis to hospital data. *Biometrics*, **2**, 47-53.
41 Hernan, M.A., Clayton, D. and Keiding, N. (2011) The Simpson's paradox unraveled. *Int J Epidemiol*, **40**, 780-785.

Tables

Table 1. Association between 11 insulin secretion SNPs and BMI (SD) in the UK Biobank study under 5 scenarios. Corrected statistics are those after correcting for index event bias using all 119,688 individuals. Note that the numbers of individuals in the "all" analyses differ slightly because people of uncertain diabetes diagnosis were excluded from the "All adjusted for type 2 diabetes" analysis.

	All (N=119,688)		Controls (N=113,203)		Cases (N=4,003)		All adjusted for type 2 diabetes (N=117,206)		All - Corrected statistics (N=119,688)	
Variant in or near gene:	**BETA**	**P**	**BETA**	**P**	**BETA**	**P**	**BETA**	**P**	**BETA**	**P**
TCF7L2	-0.024	1E-7	-0.033	8E-13	-0.118	1E-7	-0.036	1E-15	-0.0101	0.04
THADA	-0.011	0.1	-0.012	0.08	-0.060	0.1	-0.013	0.05	-0.0020	0.8
CDKN2AB	-0.008	0.2	-0.014	0.01	-0.018	0.5	-0.014	0.009	-0.0013	0.8
SLC30A8	-0.005	0.3	-0.009	0.05	-0.034	0.2	-0.009	0.03	-0.0008	0.9
CDKAL1	-0.012	0.01	-0.016	5E-4	-0.048	0.03	-0.018	1E-4	-0.0062	0.2
MTNR1B	0.011	0.02	0.008	0.08	0.004	0.9	0.008	0.08	0.0162	0.001
HHEX	-0.005	0.2	-0.007	0.09	-0.054	0.01	-0.009	0.04	0.0005	0.9
GCK	-0.005	0.3	-0.009	0.1	-0.009	0.7	-0.009	0.1	-0.0005	0.9
PROX1	-0.001	0.8	-0.002	0.6	-0.002	0.9	-0.002	0.6	0.0023	0.6
ADCY5	-0.002	0.7	-0.006	0.2	0.012	0.6	-0.005	0.3	-0.0014	0.8
DGKB	-0.003	0.5	-0.005	0.2	-0.008	0.7	-0.005	0.2	0.0021	0.6

Table 2. Examples of index event bias observed in real data when using multiple variants in genetic risk scores. BMI was inverse-normal transformed.

		UK Biobank study					Meta-analysis of independent studies				
Model	**Samples**	**BETA**	**LCI**	**UCI**	**P**	**N**	**BETA**	**LCI**	**UCI**	**P**	**N**
BMI ~ insulin secretion GRS	All individuals	-0.004	-0.010	0.001	0.14	119,688	0.001	-0.010	0.013	0.8	30,440
	Type 2 diabetes controls	-0.010	-0.016	-0.005	5E-4	113,203	-0.008	-0.019	0.004	0.2	25,039
	Type 2 diabetes cases	-0.036	-0.065	-0.007	0.015	4,003	-0.037	-0.062	-0.012	0.004	5,396
	All individuals adjusted for type 2 diabetes status	-0.011	-0.017	-0.006	1E-4	117,206	-0.013	-0.021	-0.005	0.002	30,435
BMI ~ blood pressure GRS	All individuals	-0.014	-0.020	-0.008	1E-6	119,688	-0.008	-0.020	0.004	0.2	25,059
	Normotensive controls	-0.034	-0.043	-0.026	2E-16	53,377	-0.014	-0.028	0.000	0.06	18,590
	Hypertensive cases	-0.031	-0.039	-0.024	4E-16	65,584	-0.042	-0.063	-0.021	9E-5	8,267
	All individuals adjusted for hypertension status	-0.033	-0.038	-0.027	7E-31	118,961	-0.022	-0.034	-0.010	4E-4	25,049

Table 3. Association between 25 blood pressure SNPs and BMI (SD) in the UK Biobank study under 5 scenarios. Corrected statistics are those after correcting for index event bias using all 119,688 individuals. Note that the numbers of individuals in the "all" analyses differ slightly because people of uncertain hypertension diagnosis were excluded from the "All adjusted for hypertension" analysis.

	All (N=119,688)		Controls (N= 53,377)		Cases (N=65,584)		All adjusted for hypertension (N=118,961)		All - Corrected statistics (N=119,688)	
Variant in or near gene:	**BETA**	**P**	**BETA**	**P**	**BETA**	**P**	**BETA**	**P**	**BETA**	**P**
ADM	0.001	0.9	0.005	0.6	-0.012	0.1	-0.005	0.4	0.0009	0.9
ATP2B1	-0.002	0.7	-0.012	0.1	-0.011	0.1	-0.012	0.03	-0.0038	0.5
BAT2	-0.013	2E-3	-0.014	0.02	-0.016	3E-3	-0.015	2E-4	-0.0135	0.002
C10orf107	-0.008	0.1	-0.012	0.1	-0.019	9E-3	-0.016	3E-3	-0.0095	0.08
CACNB2(3')	-0.01	0.03	-0.02	1E-3	-0.013	0.02	-0.016	1E-4	-0.0100	0.02
CACNB2(5')	-0.012	4E-3	-0.023	1E-4	-0.01	0.06	-0.016	8E-5	-0.0122	0.003
CYP17A1	-0.028	3E-4	-0.044	4E-5	-0.032	2E-3	-0.038	5E-7	-0.0282	0.0003
CYP1A1	-0.011	0.01	-0.017	9E-3	-0.016	7E-3	-0.016	2E-4	-0.0110	0.01
EBF1	-0.001	0.89	-0.012	0.04	-0.002	0.7	-0.007	0.1	-0.0008	0.9
FGF5	-0.007	0.1	-0.023	4E-4	-0.014	0.01	-0.018	3E-5	-0.0078	0.08
FLJ32810	-0.008	0.07	-0.014	0.03	-0.016	8E-3	-0.015	6E-4	-0.0084	0.07
FURIN	0	0.98	-0.008	0.2	-0.008	0.2	-0.008	0.07	-0.0005	0.9
GNAS	-0.004	0.5	-0.007	0.5	-0.019	0.02	-0.014	0.02	-0.0042	0.5
GOSR2	-0.001	0.8	-0.011	0.2	-0.003	0.7	-0.007	0.2	-0.0012	0.8
HFE	0.002	0.8	-0.008	0.3	-0.001	0.9	-0.004	0.5	0.0026	0.6
JAG1	0	0.95	-0.005	0.4	-0.002	0.8	-0.003	0.4	-0.0005	0.9
MECOM	0.002	0.7	-0.006	0.3	0.001	0.9	-0.002	0.6	0.0016	0.7
MOV10	0.007	0.1	0	0.97	0.007	0.3	0.004	0.4	0.0079	0.1
MTHFR	-0.011	0.05	-0.018	0.02	-0.024	1E-3	-0.021	7E-5	-0.0110	0.05
NPR3	0.004	0.3	-0.001	0.9	-0.006	0.3	-0.004	0.4	0.0042	0.3
PLCE1	-0.001	0.8	0	0.98	-0.009	0.1	-0.005	0.2	-0.0017	0.7
PLEKHA7	-0.005	0.3	-0.016	0.01	0.002	0.8	-0.006	0.2	-0.0049	0.3
SH2B3	-0.009	0.03	-0.008	0.2	-0.023	2E-5	-0.016	4E-5	-0.0093	0.02
TBX5	-0.003	0.6	-0.004	0.6	-0.008	0.2	-0.006	0.2	-0.0031	0.5
ZNF652	0.006	0.1	0	0.98	0.005	0.4	0.003	0.5	0.0063	0.1

Figures

Figure 1. Apparently paradoxical gene-phenotype associations in the context of disease stratified genetic studies. We simulated genotype, continuous risk factor values and disease status in a general population sample according to our liability scale model and set the genetic effect on the risk factor (γ) to zero. We observed that the estimated effect of the "B" allele of a genetic marker on a continuous trait is negative both in cases and controls. Disease carriers also have higher trait value than controls. However, when combining the two strata the marker is – as expected – not associated with the trait. The reason for this apparent paradox is that the proportion of disease risk allele ("B") carriers is higher in the case group. Thus when merging cases into the control group the mean trait value of the BB group increases much more than it does in the other genotype groups. This concept is recognized as Simpson's paradox(41).

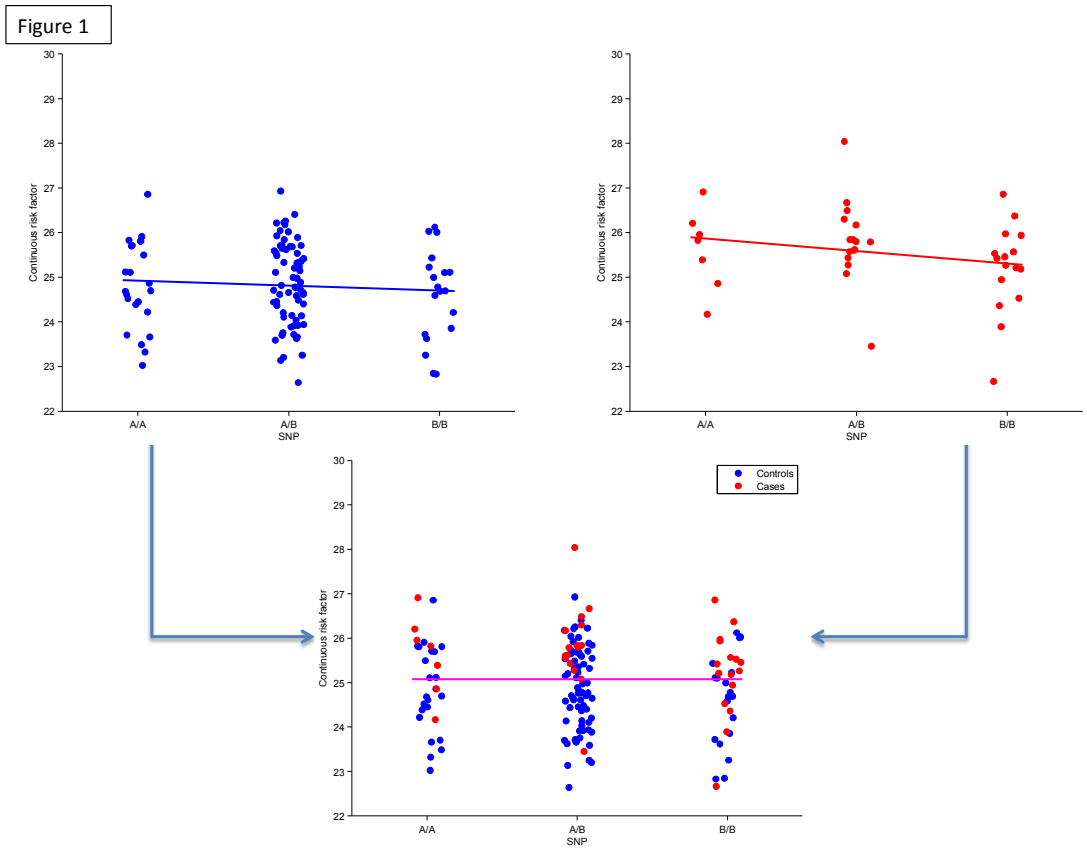

Figure 2. Enrichment for cases or controls produces spurious associations. We applied our analytical formula to compute the effect size estimate of a SNP (G) on a continuous risk factor (X) in the abovementioned liability scale model setting with the true genetic effect on the risk factor (γ) being zero. Enrichment for cases or controls produces spurious evidence of association between disease risk alleles and a risk factor correlated with the disease (equivalent to 2.5 OR per SD) in (**a**) a scenario where a risk allele (MAF 30%) increases risk with an effect equivalent to an odds ratio of 1.4 (similar to the *TCF7L2* type 2 diabetes scenario(10)) in two models: unadjusted for disease status [blue curve] and adjusted for disease status [green curve]. Dashed lines represent 95% confidence interval (CI) around the effect estimate assuming a population of 100,000 individuals. Panel (**b**) displays the same curves, but for a SNP with a rare protective allele (MAF 2%) that reduces risk of disease with an effect equivalent to an odds ratio of 0.5 (similar to the *CCND2* type 2 diabetes scenario(9)). Vertical dashed red line at 0.05 indicates the true general population disease prevalence.

Figure 2a

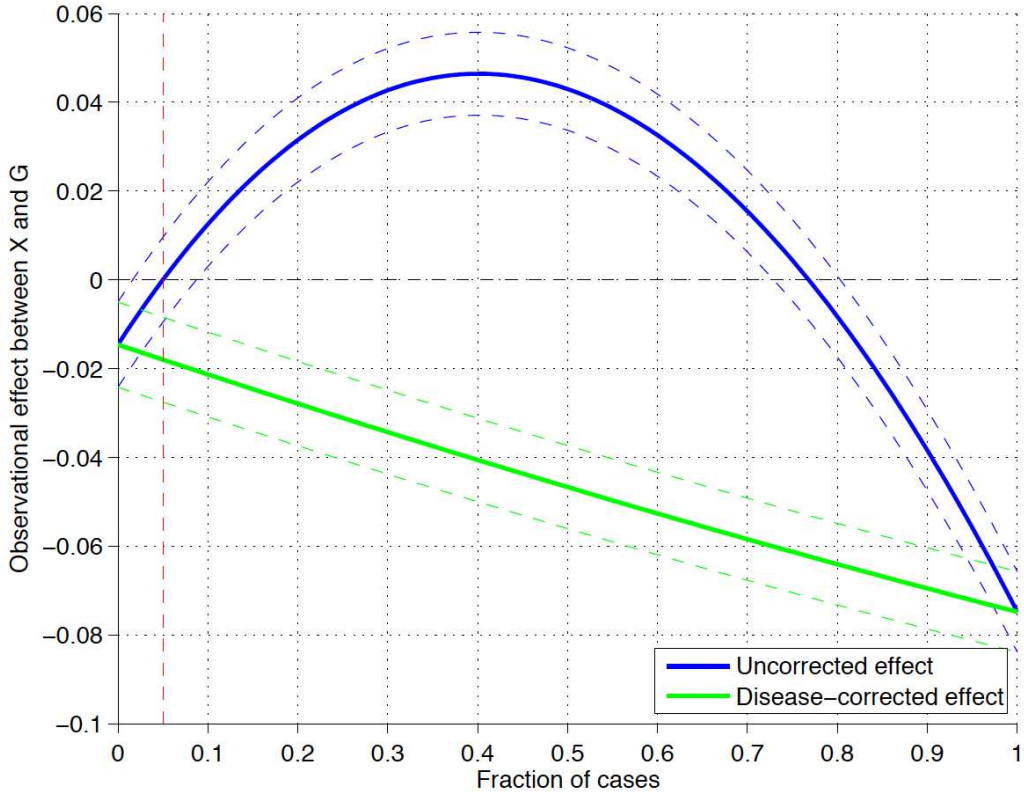

Figure 2b

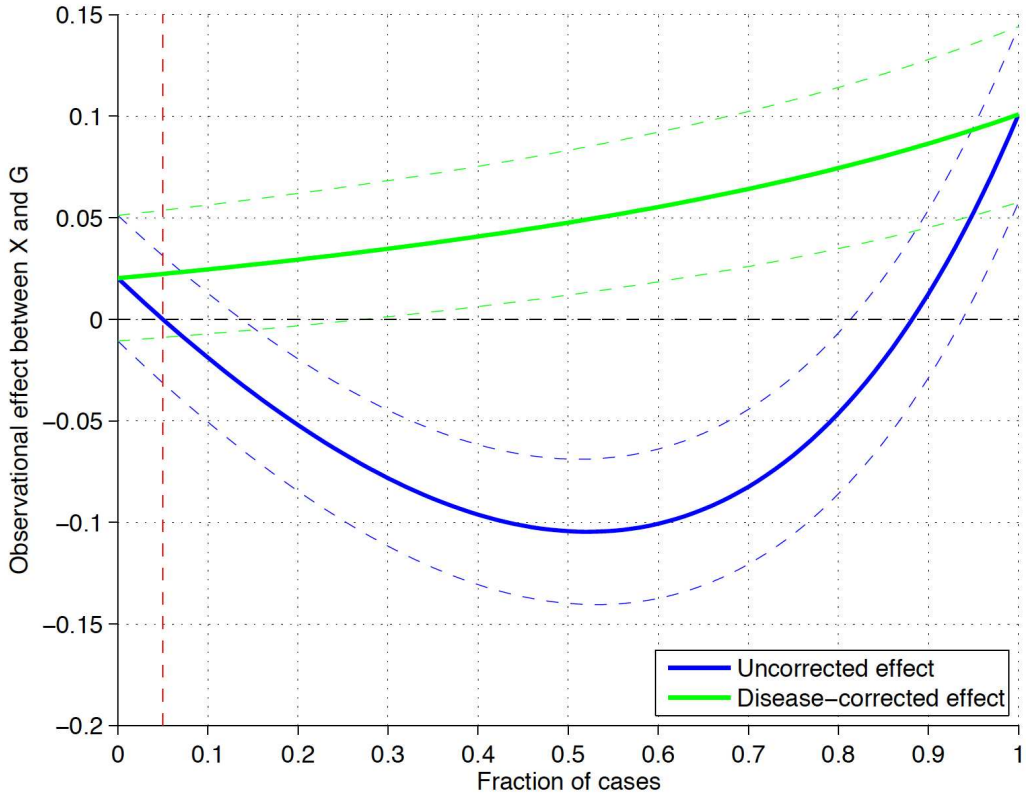

Figure 3. Scatter plot of the observed effect of type 2 diabetes-associated SNPs on BMI in the total UK Biobank sample vs. the index event bias corrected effect. The effect corrected for index event bias (shown on the y-axis) was calculated assuming the previously established 10% population prevalence of type 2 diabetes ($\pi_0 = 0.10$). Dashed line represents the identity line, where the two effects are equal. While for most SNPs the absolute value effect size estimate after IEB correction is reduced, *MTNR1B* shows increased effect size upon correction. Only this latter SNP produced a P-value surviving multiple testing correction (P<0.05/11).

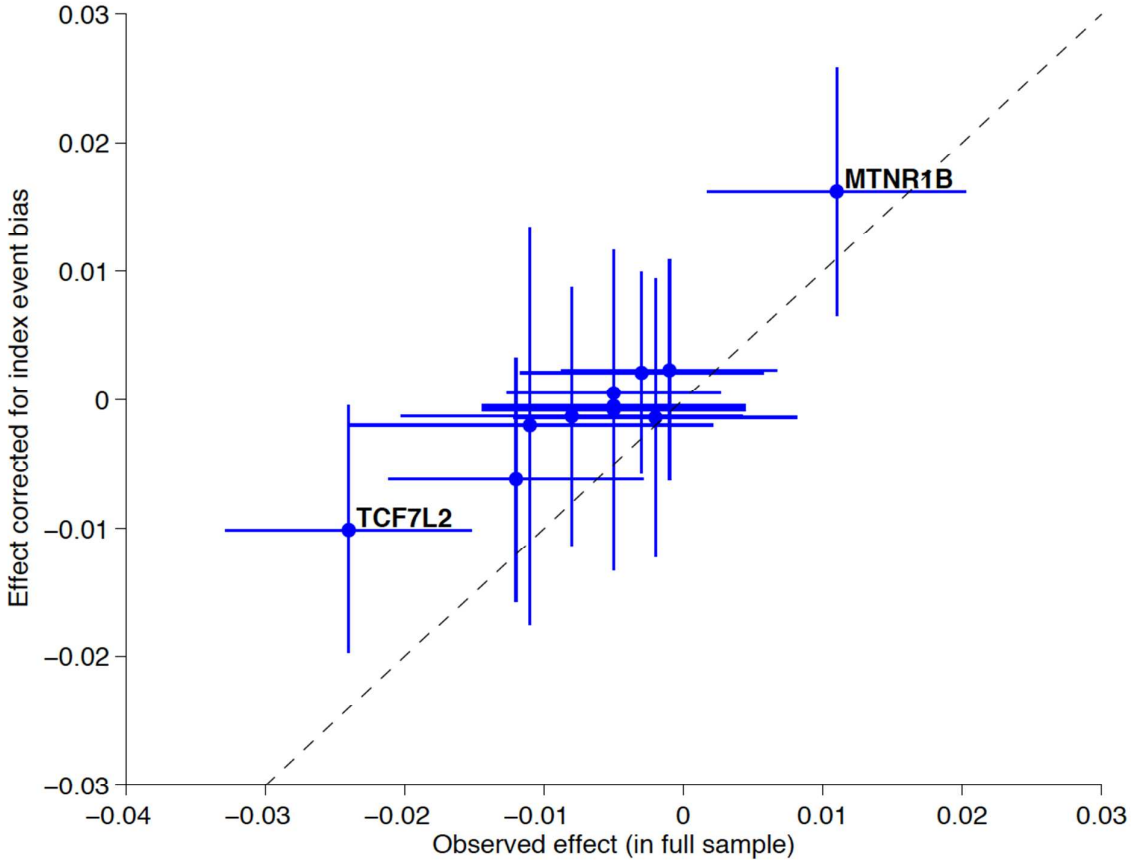

Figure 4. Index event bias in real data. (**a**) The genetic risk score (GRS) associated with higher risk of type 2 diabetes is associated with lower BMI in cases and controls separately and when combined but adjusted for type 2 diabetes status. (**b**) The GRS associated with higher risk of hypertension is associated with lower BMI in hypertensive cases and controls separately and when combined but adjusted for hypertension status. The x-axis is the effect size per disease risk allele. The vertical solid line is the null effect.

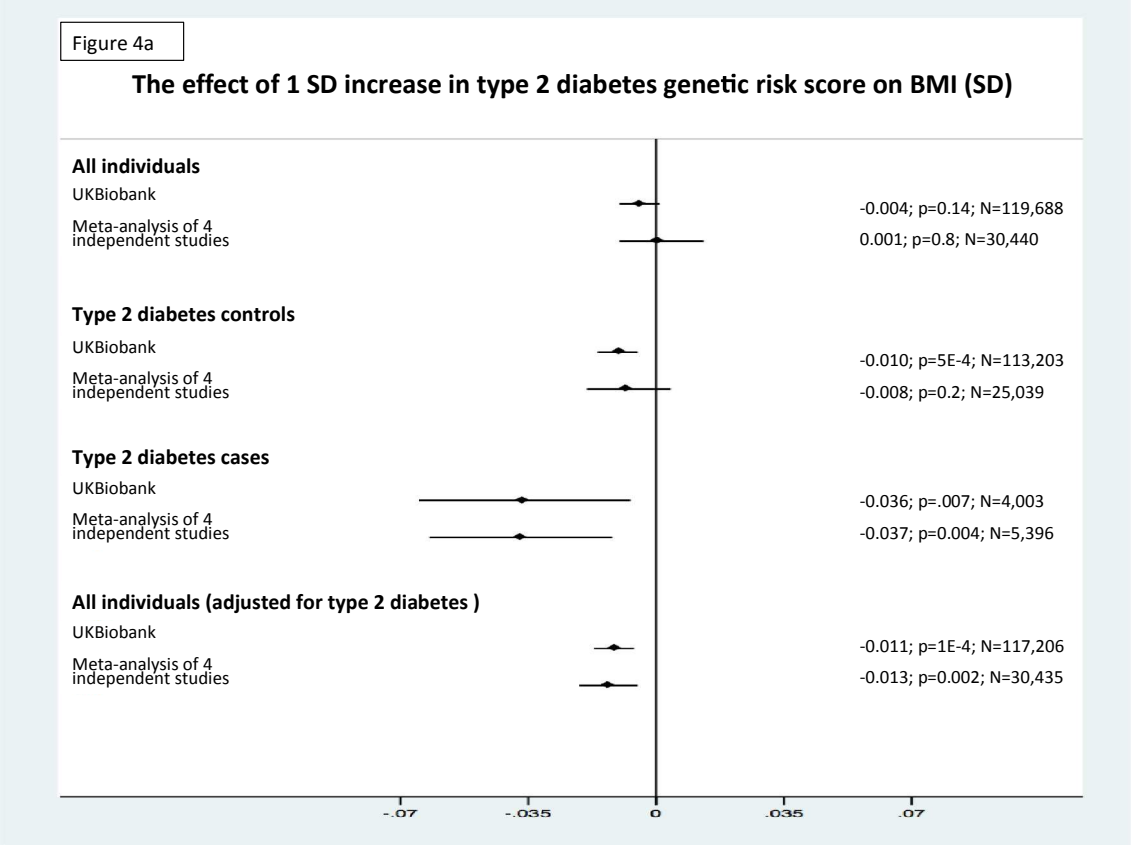

Figure 4b

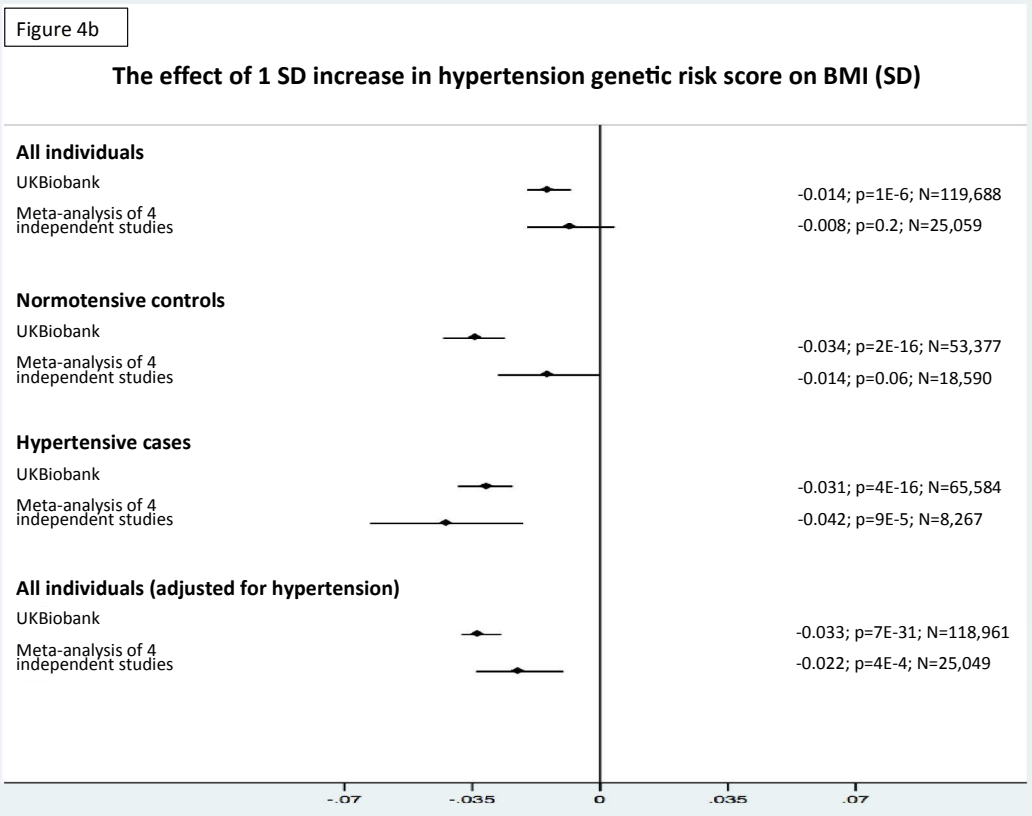

Figure 5. Power calculation to detect IEB. Using the analytical formula for IEB we derived the minimal necessary sample size to observe IEB in a study at nominal (alpha=0.05, top panels) and genome-wide significant level (alpha = 5E-8, bottom panels) with 80% power. We fixed the disease prevalence in the general population to 10%. The SNP-disease odds ratio was varied between 1 and 2.3 and the observed population prevalence of the disease was explored for the full range of 0-100%. The SNP MAF was set to 30% in the left panels and to 2% in right panel.

Figure 5